DOCUMENT RESUME

ED 090 313 TM 003 602

AUTHOR Nicolich, Mark; And Others
TITLE Demonstration of Techniques for Optimizing the Use of Criterion-Referenced Achievement Tests in Children Enrolled in a Day Care Center.
PUB DATE 74
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Achievement Gains; Achievement Tests; *Criterion Referenced Tests; Day Care Programs; Educational Diagnosis; Feedback; Instructional Programs; *Item Analysis; *Measurement Techniques; Multiple Choice Tests; Preschool Children; *Preschool Evaluation; Program Evaluation; Statistical Analysis; Summative Evaluation

ABSTRACT

This investigation describes (a) use of an item analysis technique applied to pretest results of Tests of Basic Experiences (Moss, 1971) with children 3-5 years old enrolled in a day care center; and (b) development of a new statistical technique for evaluating change ratings between pre- and posttests. A technique of obtaining mean change ratings for each item, determining significance of these ratings, and comparing item categories based on instructional objectives showed certain means for categories included in instruction significantly different from categories not included and certain mean change ratings significantly different from zero for categories of items included in instruction. (Author)

ED 090313

TM 003 002

Session 17.05

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Demonstration of Techniques for Optimizing the Use of Criterion-referenced Achievement Tests in Children Enrolled in a Day Care Center

Mark Nicolich, Lorraine Nicolich, & Jane Raph

Rutgers University

This investigation describes (a) use of an item analysis technique applied to pretest results of Tests of Basic Experiences (Moss, 1971) with children 3-5 years old enrolled in a day care center; and (b) development of a new statistical technique for evaluating change ratings between pre- and posttests.

A technique of obtaining mean change ratings for each item, determining significance of these ratings, and comparing item categories based on instructional objectives showed certain means for categories included in instruction significantly different from categories not included and certain mean change ratings significantly different from zero for categories of items included in instruction.

DEMONSTRATION OF TECHNIQUES FOR OPTIMIZING THE USE OF CRITERION-REFERENCED ACHIEVEMENT TESTS IN CHILDREN ENROLLED IN A DAY CARE CENTER

Mark Nicolich, Lorraine Nicolich, Jane Raph

Rutgers University

TM 003 602

Objectives

In developing an evaluation plan useful for both instructional and summative evaluation of teaching effectiveness at a day care center, a refinement of the gross nature of current preschool evaluation seemed imperative. An effort was made (a) to involve the use of an item analytic technique to assist teachers in using pretest results of children as a means of identifying areas of knowledge, or lack thereof, and as an aid in instructional planning; and (b) to develop a new statistical technique for evaluating item change ratings on achievement-type tests together with suggested applications for preschool evaluation.

Theoretical Framework

With recent emphasis in educational funding on accountability and on performance-related objectives, the limitations of techniques currently in use for measuring program effectiveness have become increasingly apparent. Day care is a relatively new phenomenon in the United States. Problems of accountability have arisen from lack of agreement regarding the nature of day care programs, i.e., custodial care or developmental education; generally low reliability of measurement instruments for use with young children; and questions regarding validity of the measures for demonstrating growth related to curricular objectives.

Research on effectiveness of early education experiences in the past has relied heavily on IQ tests. The inappropriateness of these tests for measuring program outcomes as well as their use in a multiracial, pluralistic society has been documented frequently (Hunt, 1961, 1964). The demonstrated fact of large initial IQ gains during the first year of school entrance and then the leveling off of such gains is also a shortcoming of traditional IQ measures (Klaus and Gray, 1968).

With respect to achievement-type tests for young children, if the measuring instrument is directly related to the instructional program, not only can individual growth be shown, but socres can also be used to assess program strengths and limitations as well as to form a basis for instructional planning for the succeeding year (Sigel, 1973; Stodolsky, in press). One might also add that, while logically, the determination of curricular objectives should clearly precede selection of evaluative approaches, in fact, the many new ventures in day care have not clearly defined their objectives. Such centers might profit from an early assessment in the school year of each child's grasp of educationally relevant concepts. These centers might profit also by observing systematically the changes each child demonstrates during enrollment in a year's program.

Techniques

The techniques to be described here are designed to be used with a multiple choice, achievement type test. For maximum benefit, it is desirable to have the test items grouped, or able to be grouped, into specific curricular areas. The techniques refer first to an item analysis as a basis for planning, and second

to the analysis of item change ratings as the basic data from which different evaluation questions may be answered.

It is of note that the statistical principles involved are basic, and represent the application of first principles to an apparently complex problem. While the techniques do not consider all possible ramifications, they do point a way for future application.

Item Analysis for Planning

An Item Analysis computer program (Nicolich, 1972) has been developed which allows each teacher to determine the specific strengths and weakness of the class or of subgroups within the class. The program will score the test, print information on the score distribution, class mean and standard deviation. The item analysis section of the program yields the percent correct for each test item, and the percent of students choosing a given response for each item. The listing of students response is presented for 3 groups: the total class, those students in the upper 27% of the test grade distribution, and those students in the lower 27% of the distribution. The teacher can then, from the percent correct, determine areas of weakness. From the responses the teacher can determine if there are particular misconceptions or general lack of knowledge in poorly answered areas. In addition, data on a particular item may indicate its appropriateness or inappropriateness for a particular group. That is, if a particular item has few correct responses, investigation of the most frequently answered distractor may show poor item construction for the particular student group under consideration. When tests are constructed for national distribution it is likely that some items will be inappropriate for a particular group of

preschool children. The assumption of similar educational experience is frequently not tenable for school aged children, but the assumption when applied to preschoolers becomes clearly untenable. Not only is there wide program variation, but various home influences, such as object uses, verbal expressions and other idiosyncrasies are far more evident in the responses of these children who have had only brief and relatively recent experience with society outside the home. In particular, the upper and lower 27% groupings can be investigated to determine if these groups have an overall need different from the total class. In a homogeneous SES group which is on the fringe of the dominant culture, (as in the data to be described here) the performance of the upper 27% can be used as a rough index of the appropriateness of the items for the total group in general.

The information can then be used to help the teacher to plan the year's curriculum to best benefit the students. It is not meant to tell the teacher which particular individual items to teach to allow a large "gain score" for that year.

A frequent deficiency of preschool achievement tests is a lack of alternate forms. Thus the pre and post tests are often the same questions. Ameliorating this deficiency in the testing procedure is lack of feedback to the students as to which responses were correct, and the relatively long period between testings. However, until alternate forms are available, it is an important (if obvious) administrative detail to avoid teaching with particular items in mind.

Analysis of Item Change Ratings

Item change ratings. In conceptualizing the effect of instructional programs in achievement terms, one could say that the purpose is that each child

add to his knowledge in certain areas. With respect to any element of knowledge (sampled by a test item) it seems reasonable to credit the item with one point (+1) if an item that was answered incorrectly on the pretest is answered correctly on the posttest, to debit one point (-1) if the reverse holds, and to give the item no credit (0) for the items on which the child exhibits no change on the two test adminstrations. From pre- and posttest data an item change rating (+1,0,-1) is assigned to each item for each subject. An average change rating over all subjects in a given class can then be determined for each item. The mean change rating for each item then becomes the basic data for statistical analysis. If we keep in mind the caveat of Sigel (1973, p.108) to avoid "premature foreclosure (of data analysis) by formal statistical analysis," it is possible to scan the mean ratings, inspect inconsistencies, and examine relationships which might lead to the generating of what Sigel terms "more realistically complex statements of hypotheses." Such informal inspection of data sets can suggest the type of analysis most appropriate for a given evaluation goal. Described herein are several approaches to utilizing the item change ratings, each related to a different evaluation question.

Evaluation of significance of item change ratings. To investigate the statistical significance of item change ratings it will be assumed that random guessing applies and that each response is equally likely for each item. In what follows, it is assumed that there are four responses for each item. (If there are other than four responses, or the responses are not considered equally likely the analysis could proceed along similar lines making the necessary changes in the probabilities).

The probability that an item is incorrectly answered on the pretest and correctly answered on the posttest (both being chosen at random from among four

possible) is $\frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$. Thus the probability of a +1 item rating is $\frac{3}{16}$. Similarly, the probability of a 0 or -1 rating is found to be $\frac{10}{16}$ and $\frac{3}{16}$ respectively. The mean and variance of the item change rating is 0 and 0.375.

Thus, under the null hypothesis of random choice on pretest and posttest, the distribution of item change rating is unimodal and symmetric. A form of the Camp-Meidell inequality (Rao, 1965) states:

$$P(|x-\mu| \geq \lambda\sigma) \leq \frac{4}{9\lambda^2},$$

where μ and σ are the mean and standard deviation of the distribution of x, and λ is any positive constant. For the problem at hand; the rejection region for the null hypothesis of guessing, versus the alternative hypothesis of learning taking place, at the .05 level of significance, is mean change ratings larger than $1.29/\sqrt{n}$, where n is the number of students included in the mean change rating. This may be restated as: a mean change rating larger than $1.29/\sqrt{n}$ indicates an increase in learing of test items at the .05 level of significance.

If n is large, then the Central Limit Theorem obtains, and the rejection region for the same hypotheses at the 5% level of significance is $1.00/\sqrt{n}$.

To determine if a particular item was learned (in accordance with the aforementioned schema), the mean change score for that item would be compared with $1.29/\sqrt{n}$; if the mean change score was larger the item was learned, if not the item was not learned. A computer program, Nicolich (1973), has been written to provide the appropriate statistics to test such hypotheses.

In addition to testing individual items, this technique can be used to test groups of items. The group would be comprised of items which come from a particular curriculum area. By calculating the mean change score for all items in the group, for all students and comparing this with $1.29/\sqrt{k \cdot n}$, the learning

in the curriculum area could be tested (k represents the number of items in the curriculum group considered.). Again, for large samples when the Central Limit Theorem obtains, the critical value would be $1.00/\sqrt{k \cdot n}$.

Evaluation of Teacher Goals. In addition to testing if groups of similar curriculum items are learned, it is possible to use the item change ratings to evaluate teacher instructional goals. The teacher is asked to divide test measured curriculum categories into one of three areas: to be included in classroom instruction, definitely not to be part of the instruction, and to be an ancillary part of the instruction program. Note, individual test items are not considered, only curriculum categories; this precludes the teacher instructing specifically for individual test items.

The item change scores are then averaged for each of the three areas. It is expected that the area included in instruction would have a larger mean item score than the area not included in instruction. Fisher's randomization technique using Snedecor's F (Bradley, 1968) would be used for this test. The curriculum area means could also be tested to determine if they differed significantly from zero, again using $1.29/\sqrt{n \cdot k}$ where k is the number of items included in the area.

If the mean of the curriculum group of taught items is significantly greater then for the group of not taught items, then the teacher could conclude that the goals were met. If not, then the conclusion would be either that the particular group of taught items were not learned any better than the not taught items, or that the non taught items were learned in some other setting. It would be reasonable to then test if the mean of the taught items category was significantly larger than zero.

Application in A Day Care Setting

The following represent some results from a study where the principles outlined above were applied.

Subjects

Subjects were enrolled in an inner-city day care facility. The children were in three classrooms, two preschool classes of children ages 3- and 4-years (N=31) and one kindergarten class of children age 5-years (N=17). All children were either black or Spanish-speaking.

Instruments

The Language and Mathematics subtests of the Tests of Basic Experiences (TOBE, Moss, 1971) were selected as criterion-referenced measures of the educational program planned for the children. Cazden (Buros, 1972) concluded in her review of these tests that the design and conditions of administration were probably as good as can be obtained for children this age.

Administration

The tests were administered to individual children or in groups of not more than two or three. Spanish language instructions included in the TOBE manual were utilized when appropriate by a native Spanish-speaking tester. Tests were given in the late fall and in the late spring of the same school year.

Criterion-referenced Sources

Within each subtest of the TOBE, items were designated by the authors of the TOBE as sampling relatively specific curricular areas, not merely factual information. For example, positional and contextual meaning are contained in the Language subtest. Geometrical shapes and measurement are areas tapped in the Mathematics subtest. A list of the areas of knowledge contained in the

subtests was given to each teacher who divided the list into three categories: those areas definitely included in her instructional program; those definitely not included; and those areas of an ancillary nature.

Results

In applying these techniques to the performance of children on the TOBE in the three classrooms, it was found that in one class a Fisher's Randomization Technique analysis indicated that the means for items in the category included in instruction, according to the teacher, were significantly different at alpha .05 from the means for items in the category not included.

In testing whether any of the three categories of items had means different from zero, results indicated that for each class on each test, the category of items designated as definitely included in instruction had mean change ratings different from zero at alpha of .05. In only one class did the category of items not included in instruction achieve change different from zero. This exception was attributed to greater flexibility in that particular class where opportunities for independent learning may have been maximized by the teacher. The category of ancillary items had means different from zero (again at alpha .05) in half of the subjects tested.

For each group of children in each class on each test there was a clearly indicated group of items indicating significant change which was amenable to interpretation with relation to the curriculum.

References

Bradley, J. V. Distribution-free statistical tests. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Cazden, C. Review of Tests of Basic Experience. In. O. Buros (Ed.), The seventh mental measurement yearbook. Highland Park, New Jersey: Gryphon Press, 1972.

Hunt, J. McV. Intelligence and experience. New York: Ronald Press, 1961.

Hunt, J. McV. The psychological basis for using preschool enrichment as an antidote for cultural deprivation. Merrill-Palmer Quarterly, 1964, 10, 209-248.

Klaus, R. A., & Gray, S. W. The early training project for disadvantaged children: A report after five years. Monographs of the Society for Research in Child Development, 1968, 33 (4, Whole No. 120).

Moss, M. H. Examiner's Manual for Tests of Basic Experiences. Monterey, California: CTB/McGraw-Hill, 1971.

Nicolich, Mark J. An Item Analysis Program for Multiple Choice Examinations. Available from the author for the cost of reproduction and mailing. (1972).

Nicolich, Mark J. Item Change Analysis Program for Pre-Post Multiple Choice Examinations. Available from the author for the cost of reproduction and mailing. (1973).

Rao, C. R. Linear Statistical Inference and its Applications. New York, N. Y.: Wiley and Sons, 1965.

Sigel, I. E. Where is preschool education going? Assessment in a Pluralistic Society: Proceedings of the 1972 Invitational Conference on Testing Problems. 1973. Pp. 99-116.

Stodolsky, S. S. Defining treatment and outcome in early childhood education. In H. Walberg (Ed.), Rethinking Urban Education. New York: Jossey-Bass, in press.